Les cahiers du SIDOS

Cahier No 2

Edition et archivage de données de recherche

Au-delà de l'activité technique, une organisation, une division du travail et une définition de rôles

Reto Hadorn

Neuchâtel, 2003

SIDOS
Ruelle Vaucher 13
CH – 2000 Switzerland
Tel ++41 32 721 18 21
Fax ++41 32 721 20 74
http://www.sidos.ch/

ISSN 1424-974X

Résumé

L'archivage de données de recherche vise à faciliter l'échange de données entre chercheurs, de manière à permettre l'analyse secondaire, la réanalyse ou la consultation des sources exploitées. Il est donc l'instrument d'un enrichissement de l'activité scientifique et d'une meilleure exploitation des ressources financières mobilisées pour le recueil de données.

En tant qu'archive de données, le SIDOS entend avant tout **promouvoir une méthodologie et une organisation de l'archivage de données** propres à assurer à long terme le retour aux données originales. Le SIDOS est certes en mesure d'accueillir des données dans ses fonds, mais tient également à soutenir les efforts d'instituts qui préfèrent conserver eux-mêmes leurs données.

Le présent document présente un survol des problématiques auxquelles doit réfléchir un institut qui s'engage dans l'archivage de ses propres données. Il montrera l'importance, au-delà des choix proprement techniques, d'une division du travail, elle-même validée par une politique d'accès aux données. L'inscription de l'archivage des données dans l'organisation de l'institut représente un saut qualitatif par rapport à la pratique courante, qui relève plus du 'non effacement' que de la conservation des fichiers concernés. Au terme de ce parcours, le dépôt des données dans un service autonome dont la diffusion des données est la raison d'être apparaîtra comme le moyen le plus sûr pour assurer dans la durée l'accès aux données existantes.

Quelques enjeux de l'archivage méthodique des données

Le SIDOS a maintenant quelques années d'expérience de l'archivage de données de recherche. Les routines sont en place, les techniques rodées, la sécurité à long terme assurée, la documentation mise en forme. Un catalogue permet à l'utilisateur potentiel d'identifier les données pertinentes pour lui.

Au chercheur qui dépose occasionnellement un jeu de données, les activités du SIDOS peuvent paraître comme essentiellement techniques. Lorsqu'il effectue lui-même une sauvegarde de ses fichiers les plus actuels sur un CD-ROM, il accomplit une tâche technique.

Pour mettre en place le dispositif d'archivage du SIDOS, il a effectivement fallu répondre à des exigences techniques: définir des standards, des procédures, construire une base de données pour la description des données et le suivi de l'archivage, installer et dompter machines et programmes. Les collaborateurs sont supposés avoir une bonne maîtrise des techniques d'analyse de données et une expérience étendue de l'informatique, de manière à être à l'aise dans des activités techniques de toutes sortes.

Ces techniques sont généralement maîtrisées par les collaborateurs des instituts de recherche aussi; d'un point de vue strictement technique, ils seraient donc en mesure de procéder eux-mêmes à un archivage sûr de leurs données. C'est en tout cas ce qu'a dû se dire cet institut de recherche, qui a un jour approché le SIDOS pour savoir ce qu'il fallait précisément faire pour archiver ses données selon les règles de l'art et s'assurer de pouvoir

réexploiter ses propres données dans le futur, par exemple dans le cadre de nouveaux mandats de recherche.

C'est bien volontiers que le SIDOS a répondu à la demande. Son rôle n'est pas seulement d'archiver des données dans ses propres fonds mais aussi – et peut-être d'abord – de promouvoir la réutilisation des données existantes et donc d'en faciliter l'accès. De pouvoir compter sur des instituts de recherche sensibles à ces enjeux est pour lui de première importance.

L'expérience a été enrichissante. Alors que la question de départ était en substance de savoir quels sont les choix techniques appropriés, la réponse a mis beaucoup plus de poids sur le rôle joué dans l'équipe de recherche par les personnes chargées de l'archivage et la nécessité de valider ce rôle par une politique à long terme au niveau de l'institut lui-même. C'est sur cet itinéraire, qui va de la technique d'archivage à la politique de conservation des données que conduisent les pages qui suivent.

On verra qu'il y a, entre la vision purement technique de l'archivage et sa pratique méthodique un **saut qualitatif**. Qu'une politique de diffusion des données est indissociable d'une politique de la recherche, ainsi que de pratiques de la recherche qui donnent plus de poids aux données et au travail intellectuel qui s'y trouve investi. Que l'édition des métadonnées doit être intégrée au processus de recherche plutôt que d'être reléguée à la toute fin du projet.

Voyons donc à quoi engage un archivage méthodique des données.

Du non effacement des fichiers à l'enregistrement finalisé

Le fait de ne pas effacer les fichiers de données et de ne pas jeter le matériel imprimé relatif à une recherche passe parfois pour une «conservation des données». Cette **conservation par défaut** est fragile, parce qu'aucun travail d'édition de la documentation ne vient faciliter son approche et parce que personne n'est en charge de la conservation à long terme du matériel informatisé. Il arrive que les données soient «oubliées» sur des supports qui ne sont pas entretenus, les papiers empilés dans un autre lieu ou rangés dans des boîtes d'archives étiquetées autrement que les fichiers ou les supports magnétiques des fichiers. Si on parvient à relire les fichiers de données, il est parfois impossible de remettre la main sur la définition des variables…

L'archivage méthodique repose sur un **archivage parallèle des données et de la documentation.** L'inscription du jeu de données dans un **catalogue** comprenant des informations détaillées sur les différents objets physiques et virtuels concernés permettra de reconstituer un jeu de données complet, incluant données et documentation, même si différents objets doivent, pour des raisons pratiques, être conservés en des lieux différents.

Des données «jetables» aux données capitalisées

C'est souvent plus par ses analyses que se profile le chercheur en sciences sociales, que par les données qu'il construit. Les fichiers informatiques deviennent inutiles une fois les publications achevées. Les données, même si elles ne sont pas effacées, sont traitées comme si elles ne devaient plus jamais être utilisées.

L'archivage méthodique des données est orienté vers **l'échange des données** et la mise en valeur du travail intellectuel investi dans leur élaboration. Les données, accessibles, documentent le travail intellectuel effectué pour les construire. Un accès facilité enrichit le débat scientifique en permettant la vérification, la réanalyse, l'analyse secondaire, le prolongement de l'analyse initiale. L'accès aux données est aussi nécessaire au développement de la connaissance que l'accès à l'interprétation de ces données, aux analyses publiées. Prendre les mesures nécessaires pour assurer un accès durable aux données est un service rendu par le producteur de données à la communauté scientifique et, plus largement, au public qui soutient financièrement la recherche.

Des données privées aux données publiées

A qui appartiennent les données? Il n'est pas question de trancher ici le débat en cours, mais seulement de constater que les attitudes des chercheurs et des institutions varient grandement à cet égard. Que l'auteur d'un jeu de données en ait la **propriété intellectuelle**, tout le monde est d'accord là-dessus. La question est plutôt de savoir dans quelle mesure cela lui donne le droit d'en restreindre l'accès à long terme ou… de les oublier dans un coin. La question est également posée aux «producteurs» des données, c'est-à-dire aux organismes qui subventionnent la recherche ou décernent des mandats.

Le choix de procéder à un archivage méthodique est clairement l'expression d'une politique d'ouverture, d'accès facilité aux données et d'utilisation à long terme. L'archivage est alors un acte de publication, analogue à la publication d'un article : il suppose d'ailleurs un **travail d'édition**, propre à faire du jeu de données un ensemble données-documentation strictement coordonné, cohérent, logique. Après tout, la rédaction d'un article est aussi parcourue du souci d'être compris des lecteurs idéaux auxquels il est destiné.

Des «données» au «jeu de données»

Le chercheur entretient avec les données qu'il construit un **rapport quasi intime**. La formulation du projet, l'élaboration des hypothèses, la rédaction de l'instrument de collecte, l'expérience du terrain, le traitement des erreurs de saisie, tout cela tend à créer pour le chercheur un rapport *immédiat* aux données, un rapport dans lequel aucune médiation n'est nécessaire.

Par contre, un autre chercheur qui veut utiliser ces données a besoin d'une **médiation**, en l'espèce d'une documentation qui décrit les variables et la récolte des données de manière rigoureuse et donne de surcroît des informations sur le projet, son déroulement et le contexte dans lequel il a pris place. L'archivage méthodique est par principe réalisé *du point de vue de l'utilisateur potentiel*.

Les **utilisateurs** de données ne sont pas nécessairement des chercheurs externes à l'institut; il peut s'agir de nouveaux collaborateurs de l'institut, voire des auteurs des données eux-mêmes, s'ils veulent réutiliser leurs propres données quelques années plus tard. Au moment de la production des données, la tentation est grande de croire qu'on se souviendra *toujours* de ces détails que l'on sait déterminants pour une exploitation pertinente.

Au SIDOS, nous avons pris l'habitude de désigner l'ensemble données – documentation par l'expression «jeu de données», afin de ne pas entretenir

l'illusion que les données se suffisent à elles-mêmes.

De l'auteur des données à l'éditeur-archiviste

Parce que l'archivage des données est supposé servir des tiers, non familiers de ces données, le chercheur auteur des données n'est pas toujours dans la position la plus favorable pour procéder seul à la préparation du dossier. Sa **connaissance intime du projet** le dispense largement d'utiliser la documentation qui sera nécessaire à un tiers.

L'archivage méthodique des données est aussi une question de **point de vue**. Un archivage approprié fait appel à des compétences spécifiques et devrait être la tâche d'une personne, d'un poste, d'un service spécialement mandaté, même s'il est effectué au sein de l'institut producteur des données. Cette instance doit être en mesure de **faire valoir auprès des auteurs des données le point de vue, les besoins et les intérêts des futurs utilisateurs des données...** donc les intérêts de l'auteur lui-même, si quelques années plus tard il se trouve en situation d'aborder ses propres données. Cette instance doit aussi être en mesure d'exprimer des **exigences** à l'égard de l'auteur des données.

Vu sous l'angle de la communication des données du producteur-auteur à l'utilisateur-lecteur, l'archiviste est dans la position de l'éditeur responsable.

Du jeu de données à la collection

Le chercheur ou l'équipe de recherche traite un jeu de données, au mieux un petit nombre de jeux de données en parallèle. La conservation de ces données se présente à chaque fois comme un **cas particulier**, pour ne pas dire un cas unique. Cette situation est peu propice au développement d'une méthodologie, de routines, d'un véritable concept d'archivage.

L'archive éditrice de jeux de données travaille au niveau de la *collection.* Elle doit passer d'une approche individualisée des jeux de données à une approche méthodique, organisée avec système. Cette approche nécessite la définition de **standards** appropriés pour la description du jeu de données, la documentation des variables et les formats de conservation. Ces standards sont un facteur de qualité lorsqu'ils concernent les contenus et un facteur de continuité lorsqu'ils se rapportent aux formats de conservation.

L'intérêt des standards ne réside pas seulement dans le caractère homogène de la collection qui en résulte, mais aussi dans le fait qu'ils demandent à être formulés et expliqués: la définition de standards est l'occasion d'expliciter la **rationalité** de l'édition, de la conservation et de la redistribution des données archivées.

De l'archivage a posteriori à l'archivage a priori

L'archivage des données est le plus souvent effectué au terme du processus de recherche. De toute évidence, il faut d'abord produire ce qui va être déposé aux archives. Cependant, si la question de l'archivage des données n'est posée qu'à ce moment-là, le jeu de données se présente parfois avec ce **profil en négatif** esquissé dans les paragraphes ci-dessus: «juste pas effacé», traité comme un instrument provisoire, subordonné à l'exploitation, considéré comme le bien privé du chercheur (de l'équipe, de l'institut) et géré comme un cas particulier. Pour se rapprocher du standard, un gros travail doit être accompli, que l'on croit alors lié à l'archivage, constitutif de l'archivage des données.

L'archivage méthodique est d'autant plus aisé qu'il est **anticipé**, que les données ne sont pas produites seulement à des fins d'analyse mais aussi de **publication.** Si l'ensemble des opérations de recherche sont faites dans la perspective d'une publication des données, le travail d'édition qui se présente au moment du dépôt du jeu de données est considérablement réduit. Même si les données ne sont déposées aux archives qu'en fin de projet, la perspective de l'archivage doit faire partie intégrante de tout le processus de recherche.

De la technique d'archivage à l'organisation d'une archive

L'archivage d'un jeu de données peut apparaître comme une activité essentiellement technique, préoccupée de formats, supports et copies de sécurité. La brochure éditée par le SIDOS dans le but d'aider les chercheurs à préparer leur matériel en vue d'un dépôt dans ses fonds tend probablement à confirmer cette impression. [1]

Et pourtant… On vient de lire ci-dessus la description d'un certain nombre de choix qu'il faut effectuer pour rendre les données accessibles et utilisables à long terme. Comme on a pu s'en rendre compte, ces choix ne vont pas tous de soi, parce qu'ils impliquent pour l'auteur des données qu'il adopte le point de vue d'un utilisateur distant dans l'espace, le temps et donc dans l'espace culturel. Il faut également que l'on songe au traitement de collections de jeux de données plutôt que de jeux de données isolés.

Pour parcourir ce chemin, il faut plus que l'accomplissement de rites techniques; **l'archivage méthodique de données suppose la définition d'une politique (de présentation, d'archivage et d'accès aux données), une organisation du travail, une distribution des rôles.** L'institut producteur de données qui souhaite développer une pratique d'archivage interne ne peut pas se contenter d'une solution purement technique: il doit mettre en place un dispositif comprenant tous les aspects évoqués ci-dessus, politiques, organisationnels et techniques, et surtout le maintenir dans le temps: tout cela suppose une **politique d'institut**, présentant à long terme la stabilité nécessaire.

Le SIDOS est un organisme d'édition, d'archivage et de diffusion de données, qui incarne une telle politique. Sa spécificité est d'être en tant que tel l'expression de cette politique. L'institut de recherche collecteur de données qui veut archiver lui-même ses données doit élaborer pour lui-même une telle organisation et l'inscrire dans une unité d'organisation dédiée, qui viendra s'ajouter aux autres. Il y aura nécessairement concurrence pour les ressources. Nous reviendrons plus loin là-dessus.

L'édition du jeu de données

Le matériel qui fait l'objet des activités d'édition et d'archivage est déjà largement décrit dans la brochure citée plus haut. Il n'est présenté ici que sous une forme très résumée.

Un jeu de données prêt à l'archivage comprend un ou plusieurs **fichiers de données** ainsi que la **documentation des données et du processus de production des données**: les instruments utilisés pour le recueil de données, les consignes aux intervieweurs, les listes de codes, les définitions de variables, une liaison claire entre l'instrument de collecte et le fichier de données, des informations

[1] «Archivage de données; préparation des données et de la documentation». Neuchâtel, mai 1995. Cette brochure peut être commandée au SIDOS ou consultée on line à l'adresse suivante: http://www.sidos.ch/.

détaillées sur la collecte et le contrôle des données ainsi que des informations sur les objectifs du projet et son cadre théorique. Un **codebook** comprend dans l'idéal toutes les informations se rapportant aux variables, y compris des remarques méthodologiques éventuellement nécessaires. Le **rapport technique** sur la collecte de données détaille les procédures suivies, les conditions de réalisation et analyse les non réponses. On trouvera généralement dans les **publications** des indications sur le sens de la démarche de recherche, l'approche théorique et la posture idéologique des auteurs.

Les données (codes) et leur définition (sens) doivent être articulées de manière rigoureuse. Toutes les variables présentes dans le fichier doivent être décrites et renvoyer soit à une source dans l'instrument de collecte, soit à un algorithme de calcul basé sur des variables clairement définies. Lorsque la documentation décrit des variables n'existant pas dans le fichier, c'est généralement l'indice de manipulations du fichier, qui demandent à être expliquées. Les valeurs prises par les variables doivent se limiter aux valeurs définies et leur sens doit être facilement accessible (labels dans le fichier de données, listes de codes, codebook). Les codes sauvages irrécupérables sont recodés en une valeur manquante distincte des autres. Les valeurs manquantes induites par des filtres prennent une troisième valeur.

L'instance chargée de l'archivage de données effectue en principe des **contrôles** sur ces traitements, demande aux auteurs des données les informations complémentaires nécessaires, commente les problèmes qui subsistent dans le jeu de données et supprime du fichier les variables non définies.

Les dispositifs techniques

Un archivage méthodique suppose que soient prises une série de mesures, qui vont être détaillées dans les paragraphes suivants.

Le choix de formats standards

L'archive éditrice des données doit baser son travail sur un choix limité de formats traités comme des standards.

Longtemps, on a pensé que ces formats devaient être très basiques (fichiers ASCII) pour faciliter les transferts entre systèmes informatiques, assurer l'indépendance à l'égard de logiciels qui imposent des formats propriétaires et réduire les conversions de formats induits par l'évolution des techniques informatiques. Depuis, l'évolution des logiciels a fait croître les exigences des utilisateurs. Si les principes de base demeurent, on tend aujourd'hui à les nuancer. Un format propriétaire est acceptable si sa définition est publique, s'il est utilisé par plusieurs logiciels distincts, si le logiciel considéré est d'utilisation courante sur diverses plates-formes informatiques ou que les filtres de conversion sont disponibles.

On peut à la limite accepter des formats propriétaires s'ils restent en nombre très limités et appartiennent à des familles de logiciels sur l'avenir desquels il est raisonnable de parier. La convertibilité entre les systèmes informatiques les plus répandus reste un critère important. Un équilibre doit être trouvé entre l'efficience du traitement (souvent plus grande avec des formats propriétaires) et la lisibilité à long terme.

La standardisation des formats doit également simplifier les opérations courantes dans la procédure d'archivage

en favorisant la mise en place de routines de traitement.

Pour les **données quantitatives**, le format SPSS portable est considéré comme un standard adéquat. Le programme lui-même est largement répandu et le format portable, écrit en pur ASCII, indépendant du système informatique. Le programme DBMS-Copy, du même producteur, permet de convertir les fichiers SPSS dans un grand nombre d'autres formats. Le système des étiquettes de variables et de valeurs permet d'intégrer une partie de la description des variables dans le fichier de données lui-même. Il est en tout temps possible de convertir un fichier SPSS en un fichier de données brutes ASCII tabulées et d'accompagner ce fichier d'une description du fichier, produite par SPSS: il est dès lors possible d'importer les données conservées dans le format SPSS portable dans n'importe quel programme d'exploitation statistique sur n'importe quel système informatique.

Les **bases de données** plus complexes, telles que les bases de données relationnelles comprenant plusieurs tables, posent plus de problèmes. La grand nombre des systèmes de gestion de bases de données oblige l'archive à se rabattre sur un format de conservation très basique. Il n'y a dans ce domaine aucun format standard, même si le format jadis créé par dBASE a une descendance importante. La solution consiste ici à créer un fichier textuel tabulé à partir de chaque table de la base de données et de conserver en parallèle une définition précise des tables, des champs et des relations entre tables.

Une base de données ne se limite généralement pas à une structure de données. Des fonctionnalités y sont intégrées, qui permettent la saisie, l'édition et l'effacement d'éléments, l'interrogation, ainsi que la production de diverses formes de rapports. Une documentation complète comprendra la liste des plus importantes de ces fonctionnalités. Certaines d'entre elles devront éventuellement être reprogrammées si les données sont chargées dans un autre système de gestion de base de données.

S'il se trouve que l'archive travaille avec le logiciel utilisé pour créer la base de données, elle peut mettre la base de données à disposition dans le format original tant qu'il est reconnu par les descendants du logiciel. La conservation à long terme n'est cependant assurée que par les fichiers textuels tabulés.

Les **données textuelles** sont de préférence conservées dans le format d'un traitement de texte courant, choisi par l'archive comme standard, en évitant les mises en forme compliquées. Une copie en format textuel brut, sans formatage, assure le long terme.

La **documentation** peut se présenter sous des formes très diverses.

- La **documentation imprimée** doit être conservée dans tous les cas sous cette forme. Les conditions d'une conservation à long terme sont cependant difficiles à réunir. On sait aujourd'hui le support papier fragile, de même que les photocopies qui sont le plus souvent destinées aux archives. Il est donc indispensable de **numériser la documentation papier** afin de la faire bénéficier des mêmes routines de conservation que les fichiers de données. Le format TIFF (mode de compression 4) passe actuellement comme le plus sûr. Pour la consultation et la diffusion, les fichiers TIFF peuvent être intégrés dans un document PDF (Portable Document Format).
- Les documents existant déjà sous forme électronique sont conservés

dans le format original et sous forme imprimée. Pour une conservation à plus long terme, il est probablement utile de les convertir au format PDF aussi. Au SIDOS, les documents papier numérisés et les documents électroniques sont fusionnés dans un document PDF unique, utilisé notamment pour la transmission de la documentation à l'utilisateur des données.

Le programme le plus courant pour le traitement des fichiers PDF est la suite Acrobat de Adobe. Bien que propriétaire, ce format est considéré comme relativement sûr, d'une part parce que sa définition est publique, donc l'écriture de programmes de conversion toujours possible, d'autre part parce que plusieurs producteurs indépendants proposent des logiciels en mesure de manipuler le format. Les archives de données les plus avancées dans la digitalisation de la documentation prennent toutes cette option.

A noter l'existence d'une stratégie complémentaire, qui consiste à imprimer tous les documents (même ceux qui parviennent à l'archive sur support électronique) sur un papier sans acide, dont la conservation est garantie sur une durée de 500 ans.

Le choix d'un dispositif de conservation

Le dispositif de conservation est une combinaison appropriée de supports informatiques, de procédures de copie et d'entreposage en des lieux multiples ainsi que de rafraîchissement des enregistrements électroniques. Les solutions concrètes sont diverses selon la taille de l'institut et le volume de données à conserver.

Les points cruciaux sont les suivants:

- Le matériel à conserver doit exister en **au moins deux exemplaires, conservés en des lieux différents**, l'un d'eux étant externe à l'institut.
- En cas de recours à des supports magnétiques, il convient de relire et **réécrire ceux-ci à intervalle régulier**. L'intervalle dépend du support choisi. Les CD-ROM ont une durée de conservation plus longue, mais pas encore complètement connue et certainement beaucoup plus courte qu'on ne le prétendait au départ. Un CD-ROM de qualité, conservé dans des conditions optimales, n'est pas garanti au-delà de 50 ans. Le papier sans acide bat tous les records avec 500 ans (exception faite de la pierre taillée, conservée dans un musée qu'il faudra éventuellement reconstruire plusieurs fois…).
- Si le volume de données (ou de documentation) croît trop vite par rapport aux supports à disposition, une compression peut être envisagée - qui doit néanmoins toujours être effectuée avec le même programme, érigé en standard.

Exemples de dispositifs:

- L'archive est conservée dans un domaine réservé sur un disque dur du serveur sur lequel est effectué un back-up quotidien. Une copie est effectuée deux fois par mois, par exemple sur cassette, et conservée en un lieu externe, le domicile d'un chercheur ou le coffre d'une banque. Une cassette par année est conservée à long terme et sera recopiée tous les trois ans pour assurer la pérennité de l'information.
- Comme ci-dessus. Les jeux de données dont le traitement est achevé sont enregistrés sur au moins deux CD-ROM distincts et effacés du serveur, ce qui libère de la place pour les nouveaux

jeux de données. Les CD-ROM sont recopiés à intervalle régulier, par exemple tous les 10 ans si les conditions de conservation ne sont pas optimales. La copie peut aussi être effectuée sur un CD-ROM et une bande, de manière à réduire les risques liés à un support spécifique.

Le choix d'un dispositif de description et de catalogage

Les jeux de données archivés doivent être décrits de manière homogène, tous selon le même modèle. Il s'agit donc de choisir les rubriques pertinentes et un support adéquat pour cette information (probablement un système de fiches électroniques, voire une base de données si la collection de données est importante.).

Le dispositif de description remplit les fonctions du catalogue qu'entretiennent les bibliothèques et sert notamment à retrouver les objets conservés (système de cotes). Une description brève du projet et des méthodes de recueil de données, accompagnée d'une liste des thèmes traités par l'instrument de collecte de données, devraient permettre à l'utilisateur de sélectionner les jeux de données qui l'intéressent sans avoir à consulter la documentation complète.

Le catalogue de données publié par le SIDOS sur le serveur www.sidos..ch est un exemple d'un tel système. Il s'appuie dans sont contenu sur les standards internationaux; il est géré dans une base de données relationnelle.

L'expression la plus actuelle du standard international de documentation des jeux de données est connue sous le nom de «Data Documentation Initiative»; ce standard peut être consulté sur la page www.icpsr.umich.edu/DDI/; il inclut d'ailleurs la description des variables.

Le choix d'une organisation des objets conservés

Pour la conservation sur support électronique, il faut au minimum une organisation des répertoires sur le serveur ou une organisation des supports externes, ainsi que des conventions d'identification.

La documentation papier qui se rapporte aux jeux de données archivés doit être déposée en un lieu unique et classée selon une logique analogue.

Données et documents doivent être **entreposés de manière strictement coordonnée**. Les fichiers de données doivent renvoyer de manière univoque à un ensemble de documents clairement identifiables, organisés de manière transparente. Des cotes, enregistrées dans le système de description des jeux de données archivés et reportées sur les divers objets concernés (noms de répertoires, noms de fichier, CD-ROM), peuvent remplir cette fonction de coordination.

Banal? Allez demander à quelques instituts de recherche où sont leurs données de 1986…

Le choix de procédures d'archivage

Dans les instituts où le volume de données est important, il vaut la peine de penser au développement d'un instrument de gestion qui permette de suivre les jeux de données au cours de la procédure d'archivage. Une solution consiste à développer le système de fiches électroniques ou mieux, la base de données utilisée pour la description des jeux de données.

Organisation de l'archivage

Même s'il n'est pas un projet de recherche, l'archivage de données doit être **géré comme un projet**, c'est-à-dire être intégré aux procédures de gestion et de contrôle des activités propres à l'institut. Au même titre que l'économat, le secrétariat ou une équipe de recherche, l'archive doit être un **organe de l'institut**.

Au sein de l'institut, un collaborateur doit être explicitement et formellement mis en charge de ce projet; il sera alors en mesure d'accumuler de l'expérience, de développer des compétences et de manifester auprès de ses collègues des exigences propres à sa **fonction d'éditeur et de conservateur des données**. Outre la gestion quotidienne des procédures d'archivage, cette personne s'assurera que l'archive de données soit en mesure de survivre aux changements techniques et organisationnels qui surviennent au sein de l'institut.

Cette responsabilisation d'un collaborateur n'implique pas qu'il accomplisse lui-même toutes les opérations qui aboutissent à la conservation d'un jeu de données. Sa tâche consiste en partie à promouvoir auprès de ses collègues chercheurs cette **discipline de travail dans la documentation et le traitement des erreurs** qui fait gagner du temps au moment du dépôt du jeu de données dans les fonds. Par exemple, il encourage les chercheurs à choisir d'emblée une formulation des labels suffisamment complète qui soit rapidement comprise d'utilisateurs qui ont un contact moins immédiat avec les données; les abréviations font souvent l'effet de formules ésotériques pour qui ne parle pas la même langue ou vient d'une autre culture scientifique. Il sera attentif aux informations échangées de manière informelle dans l'équipe, qui peuvent avoir un intérêt pour un utilisateur futur: toute information utile doit être protocolée en bonne et due forme. En veillant au respect de ce type de règles, l'archiviste représente les intérêts du tiers qui, plus tard, utilisera ces données.

Dans le domaine de **l'informatique**, l'archiviste doit pouvoir recourir aux compétences d'un collègue ou d'un service informatique bien doté.

Par la force des choses, la création d'une archive de données apporte des **changements dans l'organisation du travail des chercheurs**. Des résistances peuvent se développer. Du temps sera peut-être nécessaire pour les surmonter; seule une réelle volonté de développer l'archive permettra de mobiliser les ressources nécessaires. C'est d'un processus d'apprentissage qu'il s'agit. Aussi la mise en place et le maintien d'un dispositif d'édition et d'archivage de données au sein de l'institut de recherche suppose-t-il une **politique d'archivage,** voulue par la direction de l'institut. Cette politique d'archivage s'exprimera dans des choix, des priorités et surtout dans l'encouragement à un archivage de qualité.

Les étapes de l'édition et de l'archivage

Dès la **planification** d'un projet de recherche, les chercheurs doivent se préoccuper du futur archivage des données qui seront produites. Un niveau d'exigence doit être défini pour la qualité des données et de la documentation. Si la recherche est conduite par une équipe, il convient de décider qui s'occupe de rassembler les éléments d'information qui convergeront plus tard dans une documentation des données - en d'autres termes, qui se soucie

des besoins d'information d'un futur utilisateur, externe à l'équipe.

Le fichier de données qui se prête le mieux à l'archivage est le **fichier «brut»** qui résulte de la saisie et des contrôles de plausibilité; c'est à ce moment-là que le travail de documentation des données est le plus aisé, parce que les informations pertinentes sont encore très présentes à la mémoire des chercheurs. Ce fichier a aussi l'avantage de ne pas encore être émaillé de **variables construites**. On admet qu'un nouvel utilisateur des données choisira ses propres constructions.

Il en va autrement lorsque différentes variables font partie d'une échelle ou d'un index complexe construit en cours d'analyse. Si l'ont tient à archiver un fichier plus évolué contenant de telles variables, il est indispensable de s'assurer que ces dernières soient décrites en détail - algorithme et interprétation. Les simples recodifications et les essais sans suite devraient être supprimés sans pitié au moment de l'archivage.

Contrôle de qualité

Au moment du dépôt des données aux archives, des contrôles sont effectués par l'archiviste en collaboration avec le chercheur, afin d'assurer un maximum de cohérence à l'ensemble données-documentation (voir les détails dans la brochure citée plus haut). C'est à ce moment-là que la division des rôles, quoiqu'éventuellement critique, se montre la plus importante: le point de vue externe au projet est indispensable pour traquer les inconsistances, qui posent beaucoup moins de problèmes au chercheur familier du jeu de données qu'à l'utilisateur ultérieur, pour qui elles constitueront un sérieux handicap.

Il n'est pas de produit intellectuel, manufacturé ou industriel qui réponde à un standard de qualité élevé par le seul fait qu'il ait été élaboré selon les règles de l'art. **Le niveau de qualité atteint est le produit des contrôles effectués**.

Ceci est vrai aussi pour les jeux de données. C'est pourquoi toutes les archives de données ont mis en place des procédures de contrôle qui assurent la complétude de la documentation, la correspondance étroite entre données et documentation, et la consistance de la codification.[2] Si le contrôle est effectué dans la foulée du dépôt du jeu de données, il est plus aisé de retrouver une information manquante: les chercheurs concernés sont encore là et ont l'essentiel en mémoire. **Les chercheurs eux-mêmes profitent alors du travail accompli par l'archiviste**.

A vrai dire, il n'est pas possible de soumettre tous les jeux de données aux contrôles les plus approfondis, ceux-ci prenant beaucoup de temps. Le travail ne sera parfois entrepris qu'à l'occasion d'une première commande.

Si le jeu de données n'est archivé qu'à la fin du projet, la documentation conservée se présente souvent sous une forme stratifiée, les couches se rapportant à des mutations successives du projet et des fichiers de données. L'utilisateur futur est mieux servi par une documentation qui se rapporte à l'état du fichier au moment de son archivage - il y a donc un **travail d'édition** à faire.

Il existe un **risque**, lors d'un archivage interne à l'institut, que l'on renonce à ces contrôles parce que l'auteur des données en est jugé seul responsable ou parce que la responsabilité est plus ou moins diffuse, partagée entre plusieurs personnes. Pourtant, même lors d'un archivage interne, il peut arriver que deux pages d'un quelconque document échappent à la

2 Brochure p. 13.

recopie, qui sont nécessaires pour un bon usage de la variable de pondération; on peut être tenté de sauvegarder plusieurs fichiers pour être sûr de ne rien perdre, et du coup il sera difficile de reconstituer un ensemble données/documentation cohérent. La confiance réciproque entre chercheurs dans l'institut, loin d'être une garantie de qualité, est une menace pour la qualité du produit archivé.

C'est **une question d'organisation**, de déterminer si ces contrôles sont réalisés, sur quels jeux de données ils doivent porter, qui en est responsable et qui effectue les corrections nécessaires. Ici encore, on voit que l'archivage de données ne se réduit pas à une opération purement technique et qu'il est important de définir un poste, un rôle, un service qui sera chargé de ce travail, doté des compétences techniques et de l'autorité indispensables. La simple invitation, faite à tous les chercheurs de l'institut, de déposer données et documentation en un lieu convenu, ne suffit pas à réaliser un archivage digne de ce nom.

Le dépôt

Données et documentation sont ensuite déposées dans les espaces réservés à cet effet et selon les standards définis par l'institut.[3] Une description du jeu de données est rédigée et enregistrée dans le système prévu à cet effet. Il est ainsi possible de retrouver l'essentiel de l'information sur un jeu de données sans avoir à le ressortir des archives. Si données et documentation sont conservées en des lieux et sur des modes différents, la cote associée à la description aidera à rassembler les éléments épars.

La description du jeu de données peut également être intégrée dans un catalogue de données public, par exemple dans le catalogue du SIDOS.

Protection des données (accès, diffusion)

Lors d'un archivage interne des données, il convient de déterminer aussi à qui et à quelles conditions un jeu de données peut être transmis pour une nouvelle exploitation. Une décision de cas en cas, reprise à chaque demande, est peu rationnelle. Il est recommandé de définir un petit nombre de *types* de conditions d'accès et d'attribuer chaque jeu de données à un des types pré-définis.

Cette opération est aussi l'occasion de poser explicitement la question de la publicité faite au jeu de données traité et de réfléchir à l'ouverture de l'institut à la communauté scientifique, aux possibilités d'utilisation du jeu de données dans l'enseignement, aux possibilités de collaboration avec les universités si l'institut a un statut privé, non universitaire.

La politique de diffusion d'un jeu de données doit être discutée également avec le mandant qui a éventuellement financé la recherche.

Il est aussi recommandé de faire un contrat avec les utilisateurs des données externes à l'institut, afin de s'assurer de leur engagement sur les points critiques du point de vue de l'institut. Le SIDOS, par exemple, demande aux utilisateurs de données de signer une déclaration standard, par laquelle ils s'engagent à ne pas transmettre les données à des tiers, à ne les utiliser qu'à des fins d'analyse et à citer le jeu de données dans toutes les publications pour lesquelles il est utilisé.

3 Sur la définition de ces standards, voir plus haut.

Coopération avec le SIDOS

Le SIDOS est tout disposé à faire figurer dans son catalogue des jeux de données archivés selon les règles ci-dessus, de manière à faire connaître leur existence à un large cercle d'utilisateurs potentiels. Ces jeux de données sont accompagnés de la mention «Externe». Si l'institut décide ultérieurement de transférer au SIDOS les jeux de données qu'il conserve, l'opération sera réalisable sans difficulté.

Au-delà de cette collaboration minimale, il faut bien se demander quelle est la solution institutionnelle la plus efficace, de l'intégration de l'édition et de l'archivage des données sur les activités de recherche, au sein de l'institut, ou du recours à un service spécialisé dans l'édition et la diffusion des données. En effet, l'exercice accompli dans ce document, de mettre en évidence les conditions nécessaires à un traitement approprié de données qui doivent demeurer disponibles à long terme, conduit à la question suivante: ces exigences peuvent-elles être satisfaites par un institut de recherche?

L'édition-archivage des données: une fonction ou une organisation?

Les exigences formulées dans ces pages à l'égard des instituts qui voudraient constituer des archives de données internes sont élevées. Elles ne le sont pas seulement parce que la réalisation technique des diverses opérations peut prendre du temps - moins de temps naturellement si le jeu de données a dès le début du projet été traité comme un jeu de données à publier - mais aussi parce que la réalisation de l'archivage et de la publication des données suppose que soient mises en place une **infrastructure** appropriée, une **organisation sociale** de l'activité d'archivage et de publication, une **division du travail** appropriée entre chercheurs et éditeurs-archivistes, ainsi qu'une **politique des données** régulièrement réaffirmée au niveau de l'institut.

Infrastructure

L'infrastructure peut paraître ce qu'il y a de plus simple à réaliser, puisque l'institut de recherche est nécessairement équipé pour traiter les données et la documentation. Pourtant, à long terme, ce n'est pas *un* jeu de données qui sera conservé, mais une collection, ce qui change fondamentalement les données du problème. Si l'institut souhaite garantir l'accès à ses données sur le long terme, il doit nécessairement mettre en place un système d'information qui permette de retrouver de manière rationnelle les données dont l'utilisateur a besoin à un moment donné.

Jusqu'ici, les archives de données ont mis en place des systèmes ad hoc, allant de la collection de documents textuels à des bases de données élaborées – le SIDOS a choisi cette seconde solution. Il n'y a pas pour le moment de solution toute faite: la mise en place d'un service d'édition et de distribution des données suppose nécessairement la mise en place d'un tel système, coûteuse.

Un point capital dans la mise en place d'un système de conservation des données réside dans la **coordination entre données et métadonnées**; cette coordination doit assurer que données et métadonnées n'évoluent pas de manière autonome, produisant un décalage entre le fichier de données et sa description.

Les archives de données travaillent elles-mêmes sur cette difficulté. Deux

développements importants doivent être mentionnés dans ce contexte: Nesstar et Metadater.

Nesstar est un système de mise à disposition en ligne de jeux de données documentés (www.nesstar.org). Il peut être utilisé par un service isolé ou mettre en réseau plusieurs serveurs de données, selon la configuration de l'organisation qui utilise le système. Il permet l'interrogation du catalogue de données et de la définition des variables (questions comprises), la consultation des informations détaillées sur le projet et les variables, le téléchargement de tout ou partie d'un jeu de données et même l'analyse en ligne selon des méthodes d'exploitation élémentaires (tabulations, corrélations, graphiques). Le dispositif arrive actuellement (fin 2002) à un niveau de maturité qui justifie pleinement sa mise en œuvre.

Metadater est un système de gestion des métadonnées qui est supposé soutenir le chercheur dès la conception du questionnaire pour l'enregistrement de toute espèce d'information utile dans l'exploitation des données; il doit également permettre la publication des données et des métadonnées, soit sous la forme classique d'un couple données / codebook, soit sous la forme évoquée ci-dessus du système d'accès intégré (Nesstar). Dans une version plus élaborée, le Metadater doit aussi devenir un outil pour les archives de données, les accompagnant dans leurs activités d'acquisition, de contrôle et de diffusion des données.

Le Metadater est actuellement à l'état de projet. Les premières applications utilisables ne sont pas attendues avant trois ans (fin 2005).

Nesstar et Metadater devraient apporter une amélioration dans la qualité des métadonnées mises à disposition. Leur utilisation suppose cependant aussi une compréhension de leur fonctionnement et un effort dans la mise en place: donc, un investissement spécifique.

Organisation sociale et division du travail

L'éditeur-archiviste de données se trouve dans une position intermédiaire entre un producteur (chercheur, équipe ou institut) et un utilisateur de données. On a vu plus haut qu'en tant qu'éditeur diffuseur de données, il représente les intérêts des utilisateurs de données - c'est là naturellement le fondement des contrôles effectués sur les jeux de données et la complétude de la documentation. Vis-à-vis de l'utilisateur des données, il représente les intérêts du producteur, notamment dans son droit à voir son travail reconnu (citation des sources de données en cas de publication, limitation à la circulation des données). C'est le rôle de tout éditeur, que ce soit dans le domaine littéraire, musical ou médiatique.

Une différence cependant rend nécessaire une certaine **autonomie d'action**: à la différence de l'auteur de romans, les producteurs de données ne travaille généralement pas prioritairement dans l'optique de la publication des données, mais plutôt dans l'optique de la publication d'analyses élaborées sur une exploitation statistique des données. De ce fait, l'édition des données pour diffusion peut amener un conflit sur les ressources, celles-ci devant être partagées entre l'objectif de la publication des analyses et l'objectif de la publication des données.

Cette autonomie nécessaire est de toute évidence mieux assurée par un service distinct de l'institut producteur de données.

Politique d'accès aux données

Toute organisation sociale suppose un coût: le coût de la mise en place de cette organisation et le coût de son maintien. C'est le coût consenti pour surmonter les tensions et les conflits d'intérêt autour des exigences apportées par l'éditeur de données. C'est le coût en énergie nécessaire à lutter contre l'entropie naturellement croissante de tout système.

Dans un contexte où le coût de l'édition, de l'archivage et de la diffusion des données entre en concurrence avec les ressources demandées par l'exploitation primaire des données, seule une **politique de données** explicite, légitimée par le responsable concerné et régulièrement renforcée par lui donne les garanties nécessaire d'une disponibilité prolongée des données.

C'est une politique de données de ce type qui a donné naissance au SIDOS. Elle est portée au départ par l'Académie Suisse des Sciences Humaines et Sociales, soutenue par son Conseil de Politique de la Science et partiellement reprise par le Fonds National de la Recherche Scientifique. De par la position des institutions impliquées, cette politique a nécessairement plus d'avenir que celle d'un directeur d'institut, dont la bonne volonté peut être invalidée au prochain changement de direction. De ce point de vue, le SIDOS, en tant que service qui incarne cette politique, aura en principe la possibilité d'assurer une vie plus longue aux données déposées dans ses fonds.

Le SIDOS, alternative réelle à un archivage interne aux instituts

Les ressources sociales et financières nécessaires à l'édition, à l'archivage et à la diffusion des données sont difficiles à mobiliser au sein d'un institut dont l'activité principale est la recherche et la participation au débat scientifique au travers de publications.

A contrario, on voit mieux ce que peut apporter un service d'édition et d'archivage spécialisé comme le SIDOS. S'il ne peut pas totalement décharger les instituts et les chercheurs du travail d'édition et de préparation de leur jeu de données en vue de l'archivage et de la diffusion, il les décharge du souci de la conservation à long terme et du contrôle final du jeu de données.

Acteur au sein de la communauté scientifique, le SIDOS exprime par les sollicitations adressées aux chercheurs la nécessité d'un accès plus aisé aux données; et lorsqu'il demande des éclaircissements sur des points de documentation ou la codification d'une variable, il représente les intérêts du futur utilisateur. Le SIDOS incarne l'organisation du travail nécessaire pour l'archivage méthodique des données et par sa présence dans le champ de la recherche, il exprime la nécessité d'une politique d'archivage et participe au développement d'une politique de la recherche qui valorise la publication des données.

Dans ses rapports avec les chercheurs qui déposent des données, le SIDOS est souvent amené à insister sur des aspects «bêtement» techniques tels que des formats de fichier, l'absence d'une liste de codes, le besoin d'une description plus approfondie de la méthode de collecte de données. Ces demandes d'apparence triviale expriment cependant l'ensemble des dimensions évoquées ci-dessus: une position dans le réseau de distribution des données, une division du travail dans l'édition et le contrôle, une volonté politique de prendre des assurances pour que les données soient utilisables à long terme. Le SIDOS, c'est aussi une **organisation**, l'organisation d'un ensemble de ressources techniques,

une organisation interne du travail. Par la sollicitation des chercheurs à déposer leurs données, il prend en charge la politique de données formulée par l'ASSH. Il manifeste également, par son existence, **l'exigence que les données de recherche soient plus méthodiquement et plus systématiquement archivées sous une forme qui permette une réutilisation ultérieure.**

Par son activité, le SIDOS rappelle **les exigences de la démarche scientifique**: la vérification des analyses, la possibilité de pousser celles-ci plus loin ou de combiner diverses sources de données. Il rappelle que les données, même lorsqu'elles sont apparemment épuisées par les analyses publiées, constituent un **bien capitalisable**. Il rappelle que si le producteur des données est bien le propriétaire de ce bien, au sens de la propriété intellectuelle, les données de recherche sont aussi un bien collectif, qui appartient au public qui les finance. D'autres chercheurs doivent donc y avoir accès aussi.

Il n'est pas rare que le premier à payer le coût d'un mauvais entretien de ce capital soit… le chercheur qui l'a produit. Les fichiers de données peuvent devenir illisibles par le seul fait de l'évolution des systèmes informatiques. Et des données mal documentées deviennent inutilisables pour leur auteur lui-même.

Le SIDOS tient à ce rôle de **représentant des intérêts de la communauté scientifique auprès du chercheur auteur des données** au moins autant qu'aux activités d'archivage proprement dites et à la croissance de son propre catalogue. Il est prêt à soutenir les efforts accomplis par les instituts eux-mêmes pour archiver leurs propres données de manière adéquate. Mais il doit alors aussi montrer que les exigences dépassent largement les quelques choix techniques qui assurent la préservation du matériel.

Lorsqu'on réfléchit à ce qu'il faut à un institut pour procéder lui-même à l'archivage de ses données, on redécouvre le rôle fondateur d'une véritable **politique d'archivage.** On voit mieux la nécessité d'attribuer à un collaborateur **le rôle d'archiviste** - le représentant des nouveaux utilisateurs auprès des producteurs de données. On découvre éventuellement que ce rôle est mieux rempli lorsqu'il est porté par des personnes extérieures à l'institut. En d'autre termes, c'est la **dimension sociale et politique** de l'édition et de l'archivage de données qui est mise en relief.

Le SIDOS:

- collecte et diffuse des informations sur la recherche en sciences sociales actuelle;
- tient à jour un inventaire de la recherche en sciences sociales (plusieurs milliers de descriptions de projets);
- diffuse l'inventaire sous la forme d'un inventaire on line ou d'un CD-Rom.

- facilite l'accès aux données disponibles;
- promeut l'analyse secondaire, la réanalye des données à des fins scientifiques;
- apporte un appui à la formation méthodologique.

Le principal instrument de cette politique est une archive de données de recherche, intégrée dans le réseau mondial des archives.

WWW.SIDOS.CH

SIDOS:

- erfasst Informationen über die aktuelle sozialwissenschaftliche Forschung in der Schweiz;
- unterhält eine Datenbank der sozialwissenschaftlichen Forschung in der Schweiz (mehrere tausend Projektbeschreibungen);
- veröffentlicht die Datenbank im Internet und auf CD-Rom.

- erleichtert den Zugang zu den verfügbaren Daten;
- fördert die Sekundäranalyse von Daten zu wissenschaftlichen Zwecken;
- unterstützt die methodologische Ausbildung.

Das Hauptinstrument dieser Politik ist das Datenarchiv, als Teil des weltweiten Netzes von Datenarchiven.

Le SIDOS est une fondation de l'Académie suisse des sciences humaines et sociales – ASSH

SIDOS ist eine Stiftung der Schweizerischen Akademie für Geistes- und Sozialwissenschaften – SAGW